Chapter 1
Objectives of the Secondary Analysis of Electronic Health Record Data

Sharukh Lokhandwala and Barret Rush

Take Home Messages

- Clinical medicine relies on a strong research foundation in order to build the necessary evidence base to inform best practices and improve clinical care, however, large-scale randomized controlled trials (RCTs) are expensive and sometimes unfeasible. Fortunately, there exists expansive data in the form of electronic health records (EHR).
- Data can be overwhelmingly complex or incomplete for any individual, therefore we urge multidisciplinary research teams consisting of clinicians along with data scientists to unpack the clinical semantics necessary to appropriately analyze the data.

1.1 Introduction

The healthcare industry has rapidly become computerized and digital. Most healthcare delivered in America today relies on or utilizes technology. Modern healthcare informatics generates and stores immense amounts of detailed patient and clinical process data. Very little real-world patient data have been used to further advance the field of health care. One large barrier to the utilization of these data is inaccessibility to researchers. Making these databases easier to access as well as integrating the data would allow more researchers to answer fundamental questions of clinical care.

1.2 Current Research Climate

Many treatments lack proof in their efficacy, and may, in fact, cause harm [1]. Various medical societies disseminate guidelines to assist clinician decision-making and to standardize practice; however, the evidence used to formulate these guidelines is inadequate. These guidelines are also commonly derived from RCTs with

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_1

limited patient cohorts and with extensive inclusion and exclusion criteria resulting in reduced generalizability. RCTs, the gold standard in clinical research, support only 10–20 % of medical decisions [2] and most clinical decisions have never been supported by RCTs [3]. Furthermore, it would be impossible to perform randomized trials for each of the extraordinarily large number of decisions clinicians face on a daily basis in caring for patients for numerous reasons, including constrained financial and human resources. For this reason, clinicians and investigators must learn to find clinical evidence from the droves of data that already exists: the EHR.

1.3 Power of the Electronic Health Record

Much of the work utilizing large databases in the past 25 years have relied on hospital discharge records and registry databases. Hospital discharge databases were initially created for billing purposes and lack the patient level granularity of clinically useful, accurate, and complete data to address complex research questions. Registry databases are generally mission-limited and require extensive extracurricular data collection. The future of clinical research lies in utilizing big data to improve the delivery of care to patients.

Although several commercial and non-commercial databases have been created using clinical and EHR data, their primary function has been to analyze differences in severity of illness, outcomes, and treatment costs among participating centers. Disease specific trial registries have been formulated for acute kidney injury [4], acute respiratory distress syndrome [5] and septic shock [6]. Additionally, databases such as the Dartmouth Atlas utilize Medicare claims data to track discrepancies in costs and patient outcomes across the United States [7]. While these coordinated databases contain a large number of patients, they often have a narrow scope (i.e. for severity of illness, cost, or disease specific outcomes) and lack other significant clinical data that is required to answer a wide range of research questions, thus obscuring many likely confounding variables.

For example, the APACHE Outcomes database was created by merging APACHE (Acute Physiology and Chronic Health Evaluation) [8] with Project IMPACT [9] and includes data from approximately 150,000 intensive care unit (ICU) stays since 2010 [1]. While the APACHE Outcomes database is large and has contributed significantly to the medical literature, it has incomplete physiologic and laboratory measurements, and does not include provider notes or waveform data. The Phillips eICU [10], a telemedicine intensive care support provider, contains a database of over 2 million ICU stays. While it includes provider documentation entered into the software, it lacks clinical notes and waveform data. Furthermore, databases with different primary objectives (i.e., costs, quality improvement, or research) focus on different variables and outcomes, so caution must be taken when interpreting analyses from these databases.

Since 2003, the Laboratory for Computational Physiology at the Massachusetts Institute of Technology partnered in a joint venture with Beth Israel Deaconess Medical Center and Philips Healthcare, with support from the National Institute of Biomedical Imaging and Bioinformatics (NIBIB), to develop and maintain the Medical Information Mart for Intensive Care (MIMIC) database [11]. MIMIC is a public-access database that contains comprehensive clinical data from over 60,000 inpatient ICU admissions at Beth Israel Deaconess Medical Center. The de-identified data are freely shared, and nearly 2000 investigators from 32 countries have utilized it to date. MIMIC contains physiologic and laboratory data, as well as waveform data, nurse verified numerical data, and clinician documentation. This high resolution, widely accessible, database has served to support research in critical care and assist in the development of novel decision support algorithms, and will be the prototype example for the majority of this textbook.

1.4 Pitfalls and Challenges

Clinicians and data scientists must apply the same level of academic rigor when analyzing research from clinical databases as they do with more traditional methods of clinical research. To ensure internal and external validity, researchers must determine whether the data are accurate, adjusted properly, analyzed correctly, and presented cogently [12]. With regard to quality improvement projects, which frequently utilize hospital databases, one must ensure that investigators are applying rigorous standards to the performance and reporting of their studies [13].

Despite the tremendous value that the EHR contains, many clinical investigators are hesitant to use it to its full capacity partly due to its sheer complexity and the inability to use traditional data processing methods with large datasets. As a solution to the increased complexity associated with this type of research, we suggest that investigators work in collaboration with multidisciplinary teams including data scientists, clinicians and biostatisticians. This may require a shift in financial and academic incentives so that individual research groups do not compete for funding or publication; the incentives should promote joint funding and authorship. This would allow investigators to focus on the fidelity of their work and be more willing to share their data for discovery, rather than withhold access to a dataset in an attempt to be "first" to a solution.

Some have argued that the use of large datasets may increase the frequency of so-called "p-hacking," wherein investigators search for significant results, rather than seek answers to clinically relevant questions. While it appears that p-hacking is widespread, the mean effect size attributed to p-hacking does not generally undermine the scientific consequences from large studies and meta-analyses. The use of large datasets may, in fact, reduce the likelihood of p-hacking by ensuring that researchers have suitable power to answer questions with even small effect

sizes, making the need for selective interpretation and analysis of the data to obtain significant results unnecessary. If significant discoveries are made utilizing big databases, this work can be used as a foundation for more rigorous clinical trials to confirm these findings. In the future, once comprehensive databases become more accessible to researchers, it is hoped that these resources can be used as hypothesis generating and testing ground for questions that will ultimately undergo RCT. If there is not a strong signal observed in a large preliminary retrospective study, proceeding to a resource-intensive and time-consuming RCT may not be advisable.

1.5 Conclusion

With advances in data collection and technology, investigators have access to more patient data than at any time in history. Currently, much of these data are inaccessible and underused. The ability to harness the EHR would allow for continuous learning systems, wherein patient specific data are able to feed into a population-based database and provide real-time decision support for individual patients based on data from similar patients in similar scenarios. Clinicians and patients would be able to make better decisions with those resources in place and the results would feed back into the population database [14].

The vast amount of data available to clinicians and scientists poses daunting challenges as well as a tremendous opportunity. The National Academy of Medicine has called for clinicians and researchers to create systems that "foster continuous learning, as the lessons from research and each care experience are systematically captured, assessed and translated into reliable care" [2]. To capture, assess, and translate these data, we must harness the power of the EHR to create data repositories, while also providing clinicians as well as patients with data-driven decision support tools to better treat patients at the bedside.

References

1. Celi LA, Mark RG, Stone DJ, Montgomery RA (2013) "Big data" in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med 187:1157–1160
2. Smith M, Saunders R, Stuckhardt L, McGinnis JM (2013) Best care at lower cost: the path to continuously learning health care in America. National Academies Press
3. Mills EJ, Thorlund K, Ioannidis JP (2013) Demystifying trial networks and network meta-analysis. BMJ 346:f2914
4. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A, Acute Kidney Injury N (2007) Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 11:R31
5. The Acute Respiratory Distress Syndrome Network (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 342:1301–1308
6. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, Sevransky JE, Sprung CL, Douglas IS, Jaeschke R, Osborn TM, Nunnally ME, Townsend SR, Reinhart K, Kleinpell RM, Angus DC, Deutschman CS, Machado FR, Rubenfeld GD, Webb SA, Beale RJ, Vincent JL, Moreno R, Surviving Sepsis Campaign Guidelines Committee including the Pediatric S (2013) Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 41:580–637
7. The Dartmouth Atlas of Health Care. Lebanon, NH. The Trustees of Dartmouth College 2015. Accessed 10 July 2015. Available from http://www.dartmouthatlas.org/
8. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL (2006) Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med 34:2517–2529
9. Cook SF, Visscher WA, Hobbs CL, Williams RL, Project ICIC (2002) Project IMPACT: results from a pilot validity study of a new observational database. Crit Care Med 30:2765–2770
10. eICU Program Solution. Koninklijke Philips Electronics N.V, Baltimore, MD (2012)
11. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39:952
12. Meurer S (2008) Data quality in healthcare comparative databases. MIT Information Quality Industry symposium
13. Davidoff F, Batalden P, Stevens D, Ogrinc G, Mooney SE, group Sd (2009) Publication guidelines for quality improvement studies in health care: evolution of the SQUIRE project. BMJ 338:a3152
14. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based decision support. JMIR Med Informatics 2:e13